

# Red-team results need an evidential ceiling: a comment on NIST AI 200-2

Public comment on NIST AI 200-2 (Initial Public Draft): The TEVV-Athlon Framework for Evaluating AI Systems

Bandana Kaur (hackwithher) · Submitted to NIST, 5 October 2026 · [hackwithher.co.in/research/evidential-ceiling/nist-ai-200-2/](https://hackwithher.co.in/research/evidential-ceiling/nist-ai-200-2/)

Author: Bandana Kaur, Security Researcher, APIsec Labs, Contact: [bandana@hackwithher.co.in](mailto:bandana@hackwithher.co.in).

Every AI safety evaluation result has a bounded limit on what it can show about a system’s failure rate, set by the number of trials and by how sharply the procedure separates safe from unsafe systems. This limit, called the evidential ceiling is lowest when zero violations are observed. If a system’s true violation rate is  $p$ , the chance it produces no violations in  $n$  independent trials is  $(1 - p)^n$ . At  $n = 520$ , the size of a widely used public red-team benchmark, and  $p = 1$  in 1,000, that chance is 59%. A clean result is more likely than not from a system that fails once per thousand interactions. The most a zero-count result can certify at 95% confidence is an upper bound of about  $3/n$  on the failure rate: for  $n = 520$ , about 1 in 175. This is standard binomial arithmetic; we work out its consequences for AI red teaming in “What AI Red-Team Evaluations Can and Cannot Prove” (arXiv:2607.21735).

In the worked example in the draft, Violation Frequency (the Safe Block) is summarized with descriptive statistics across User Testing and Red Teaming Events, and those statistics inform organizational decisions (Section 3.4). Section 2.4 then carries the synthesis into the AI RMF Manage function. For a rare-failure Block, the likely descriptive result is 0 violations, or 0%. Nothing in Sections 2.4, 3.4, or 4.3.2 tells the evaluator that 0% from  $n$  trials means only “below roughly  $3/n$ ,” and nothing tells the decision-maker whether  $n$  was large enough for the risk level they care about. Section 4.4 and Appendix D (Table 6) ask evaluators to report uncertainty, but give no method for the zero-count case, which is where a descriptive summary shows a point estimate with no visible uncertainty. The result is over-certification at Stage 4 (Synthesize & Interrogate): “no violations found” is read as evidence of a low failure rate the assessment could not have detected.

Our recommendation is to add the following to Section 2.4, and as a row in Appendix D, Table 6:

Statement of evidential strength. For any Block measured by counting failures (violations, harmful outputs, successful attacks), the synthesis shall report: (1) the number of trials  $n$ ,

and the effective number when trials are correlated, such as multiple generations per prompt; (2) the number of observed failures  $k$ ; (3) an exact one-sided upper confidence bound on the failure rate (for  $k = 0$  at 95%, approximately  $3/n$ ); (4) the failure rate the organization treats as acceptable, and whether  $n$  was sufficient to resolve it; and (5) where available, the rate at which the same procedure elicits failures from a system known to be unsafe (a positive control). No result shall be reported as establishing a failure rate below the bound in (3); this applies with most force when  $k = 0$ .

We also suggest the Section 3.4 example include this statement for Violation Frequency, so readers see how a red-team result, clean or not, should be written up.